T U M

I N S T I T U T F Ü R I N F O R M A T I K

Be a Robot - A Study on Everyday Activities Performed in Real and Virutal Worlds

Alexandra Kirsch

TUM-I1006
März 10

T E C H N I S C H E U N I V E R S I T Ä T M Ü N C H E N

TUM-INFO-03-I1006-0/1.-FI

Druck: Institut für Informatik der
Technischen Universität München

Be a Robot — A Study on Everyday Activities Performed in Real and Virtual Worlds

Alexandra Kirsch
Technische Universität München
Intelligent Autonomous Systems Group
Boltzmannstr. 3, D-85748 Garching
kirsch@in.tum.de

March 2010

Abstract

This report presents a user study, in which we compare the behaviour for setting and clearing the table in reality and in a simulated, computer-game-like environment. The aim was to examine the potential of using a computer-game-like simulation for user studies on cognition, in particular for robot-centred studies on human-robot interaction, but also other areas such as studies about context-specific and context-independent behaviour. A simulation allows the creation of a large number of environments at low cost and enables comparisons of behaviour in reality and simulation. In the present pilot study we have considered three points of interest: 1) the differences in user skills with the used simulation, 2) comparison of human behaviour in simulation and reality performing everyday activities, and 3) comparison of behaviour in different simulated environments.

Keywords

user study, human-robot interaction, simulation

Contents

1 Motivation

This report describes in detail a user study, in which the subjects took the role of a robot in a computer-game-like simulation of a kitchen. They performed the everyday activities setting and clearing the table in their role as a robot in the simulation and executed the same tasks by themselves in reality.

The background for this study is the development of a testbed for research on human-robot interaction. Most studies on human-robot interaction currently use the Wizard of Oz method [2], where a human controls the robot to interact with a subject. This allows interesting studies on humans, but not for robot research.

An alternative approach is to develop an autonomous robot stepwise and check for human acceptance using this robot [3]. However, these robots are mostly restricted to navigation and communication behaviour. Aspects such as planning, plan execution, joint execution of complex tasks (involving manipulation), on-line model learning and high-level intention recognition are currently not studied.

These higher-level aspects are currently only partially implemented on a handful of robots worldwide. Even with a robot that integrates the state of the art on all of these issues, there would still be the problem of safety. Robots with adequate manipulators for real-world activities and the respective size are currently in an experimental stage and not designed to fulfil safety criteria like industrial robots. Another difficult point is the speed of state-of-the-art robots. Our autonomous, simulated robot needs about five minutes to set the table for two people (only plates and cups) and our real robot is even slower. In contrast, in the present user study, all subjects needed less than a minute to set the table for two persons (including cutlery) in reality. This difference in timing doesn't allow interesting joint tasks on a planning level for a human and a robot.

High-level skills for robots are mostly developed in realistic, physical simulations of robots [7, 4, 8]. For adding the human factor into the simulation, we added a computer-game-like control to the simulation, which allows the manual control of a simulated robot. An alternative could be to model human behaviour in the simulation. This, however, is only possible for highly abstract simulations. But as we are interested in complete robot systems, we use a physical simulation, where the realistic modelling of a human would involve human motion models and exact behaviour predictions. If such a model existed, the behaviour of autonomous robots would be much more advanced as it could be used to control a robot in a natural way. Besides, a predefined model of a human can never be as unpredictable and natural as the live interaction with a human, which is provided by our simulation approach with human interaction.

In the user study, we wanted to assess the potentials and limitations of using a physical simulation with the additional option to control one of the robots for conducting human-robot experiments. But the simulation approach might also be useful for other research interests. With experiments in reality and simulation, it might be possible to differentiate abstract cognitive capabilities from behaviour influenced by personal physical capabilities. The simulation doesn't only make it possible to compare reality and simulation, but also offers different environments in which behaviour differences might be observed. In particular, the pilot study focuses on three aspects:

Usability of our simulation. Although this evaluation is specific for the simulation we are using, it is an important prerequisite for the other results. We assume that any

Table 1: Answers of participants about their affinity to computers and technical devices. The numbers in the table denote the number of participants giving the respective Likert score.

Question	1: fully agree	2: partially agree	3: don't know	4: partially disagree	5: fully disagree
I work with computers regularly.	8			1	
I work in the IT industry or are a student of informatics.	6				3
I often play computer games.		5		4	
I like to try new technical devices.	3	3	1	2	

experiment in simulation is only valuable if all subjects are equally skilled in handling the simulation and if the results don't contain too much noise caused by failures.

Comparison of human behaviour in reality and simulation. The subjects performed table setting and clearing tasks in six situations in reality and controlled a robot in the same scenarios (in a simulation of the same kitchen). We examined the speed differences, cognitive aspects with respect to used objects, and the participants' own perception of the differences.

Comparison of behaviour in two simulated worlds. We considered two additional scenarios in another kitchen in simulation for demonstrating the transferability of the results to other simulated scenarios. This aspect might also serve as an inspiration for further, more focused studies on general and environment-specific behaviour of humans.

The paper proceeds with a description of the method employed in the study, which contains a short description of the simulator and its control. We then show the results of the three aspects of the study, followed by a discussion.

2 Method

2.1 Participants

The user study was performed with nine subjects, four of them female. Five of the participants were aged between 25 and 30, two were between 31 and 40, one was between 19 and 24 years old and one participant was older than 40. Table 1 shows the answers to some questions regarding the subjects' familiarity with computers and computer games. It shows that most participants are used to working with computers and a majority is an IT professional or student. About half of the participants partially agree to play computer games often, the rest partially disagree. The interest in new technical devices is distributed more or less equally, but no one was completely disinterested in new technology.

2.2 Procedure

The study was conducted in two parts for each participant: one real-world part and one in simulation. The order of the two parts was randomised, five participants performed the real-world tasks first, the others started with the simulation part.

The trials in the real world consisted of six scenarios, three for setting and three for clearing the table (see Section 2.2.2). The tasks were given to the subjects in the form of pictures as in Figures 6 and 7 together with the written description shown below the pictures. The start configurations were provided by the experimenter. The subjects all started from a predefined starting position marked on the floor and then had to achieve the end configuration given in the task specification. The order of the scenarios was randomised.

In the first two trials, the subjects were asked to perform the tasks as they would do them normally. For the last four trials, the instructions were changed and the subjects were asked to carry only one object per hand to mimic the restrictions of a robot. Two subjects who had performed the simulation trials first, only carried one object per hand, even though they were given no restrictions. In these cases, the first four trials were the "robot-like" ones and after that the subjects were explicitly asked to perform the task as they would do them at home for the last two trials.

For preparing the subjects for the keyboard control of the simulation and as an additional measurement for their ability in computer games, the first task at the computer was playing the game Tetris (the implementation in Emacs) for two minutes. After that, they were instructed in the use of the simulation, which is described below. The subjects were only told how to control the robot in the simulation, but they were not given any details on the robot's abilities and good positions to stand for grasping. The subjects were then given five minutes to test the commands of the simulation and to get familiar with the control.

Then the trials in simulation started. The first four trials consisted in simple scenarios, where only one single object had to be moved. The order of these scenarios was randomised. In the last part, the participants had to perform the same tasks for setting and clearing the table as in the real world (in a simulation of the same kitchen) as well as two additional tasks in another kitchen. Again, the order of these trials was randomised.

Before the start of the study, we asked each participant to fill in a questionnaire asking about personal data and experience with computers and computer games (the data summarised in Section 2.1). After all trials including simulation and reality, the subjects filled in another questionnaire asking about their personal experience with the simulation and the differences they felt between simulation and reality.

2.2.1 The Simulation Environment

For the physical simulation of the environment we use the Gazebo simulation environment, which makes use of the Open Dynamics Engine (ODE) — a library for simulating physical processes. The robot we use is a B21 robot, which we also use in reality (see Figure 1) [1]. The original B21 comes without arms. Whereas in reality our robot is equipped with two Powercube arms with 6 degrees of freedom, in simulation we can choose between the Powercube arms and a pair of arms modelled after the Unimation PUMA robot (having 6 degrees of freedom) with two additional slider joints in the upper and lower arms to make them extendable and an additional joint in the shoulder to make the arms more dexterous.

There is currently no physical model and respective control available for simulating a

Figure 1: Simulation (left) of the real kitchen environment (right).

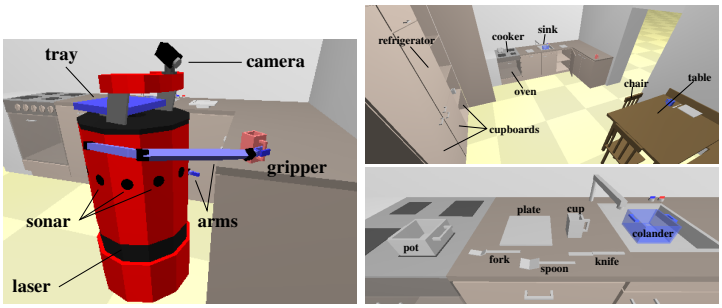


Figure 2: The robot and its kitchen environment.

human in this simulation environment. An alternative could be a non-physical simulation, but the present study is intended as a foundation for later studies on human-robot interaction, where we are interested in the implementation of the robot's behaviour. Abstracting away from the physics would make the robot research less interesting and realistic. Therefore, we opted for using the robot with the agile arms as a substitute for a human in this simulation.

The simulation offers two kitchens: Kitchen 1 is modelled after the real experimental kitchen used in the study (Figure 1). Kitchen 2 is designed after another real kitchen (Figure 2). In both kitchens there is a table, several cupboards, a sink and different objects an agent can grip: plates, cups, forks, knives and spoons (other objects were not used for this user study).

Figure 3 shows the interface of the simulation and its control. The top right window shows the outside view of the world. The small window to the left of the world view shows the local view of the robot. The window in the bottom right corner is the additional user interface for controlling the robot. It usually is an empty window, which must be active for controlling the robot.

In certain situations, additional windows of the control GUI can open like the one shown on the left to the control window in Figure 3. This window opens, when the button "E" is pressed (for "Entities") and it lists all available objects that the robot can grip. By choosing an object from the list either by mouse click or with the keyboard, the gripping process is started. Before the robot acts, it asks the user which arm it should use unless one arm is already holding an object.

For putting down objects, the user has to press the button "D" (for "put Down"). Then another window appears as shown in Figure 4, where several predefined areas are offered to the user. The exact position of each object within these areas are chosen by the robot

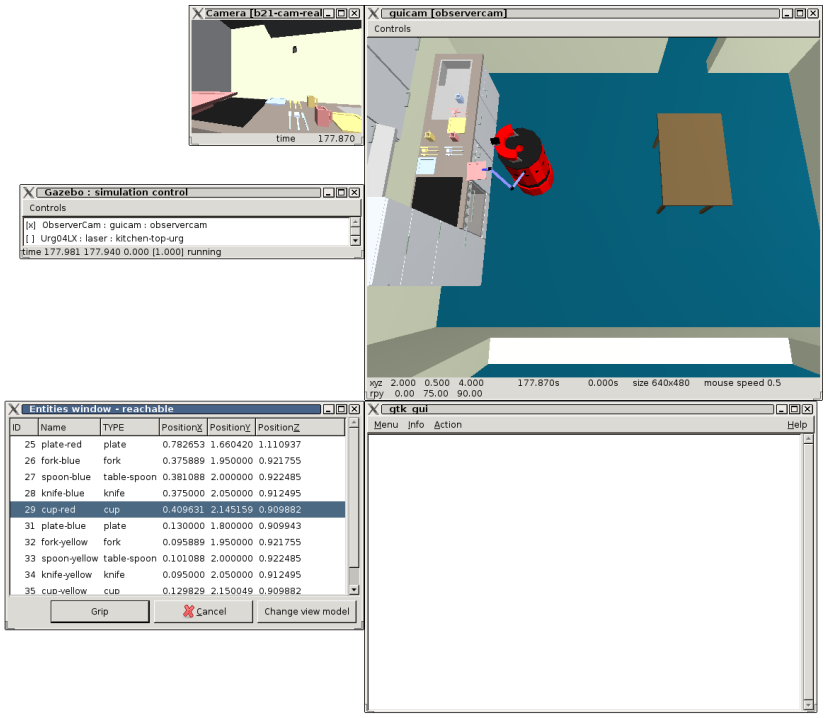


Figure 3: The simulation windows and the user interface.

automatically.

Both for gripping and putting down, the user first needs to move the robot to a position from where it can perform the manipulation task. Only the arm movement is performed automatically. If the robot is unable to perform the desired manipulation action, it responds with a message that the action is impossible.

The movement of the robot in the kitchen is controlled by the arrow keys. In the implementation used in this study, pressing an arrow key started the robot's movement (forward/backward or turning). The robot had to be stopped explicitly by pressing the Space key. Stopping only the rotation, but maintaining the forward or backward movement was possible by pressing the key "V".

The simulation we use is one of the best realistic simulation frameworks available and the ODE library for simulating physics is a widely used tool. Still, some processes in the simulation can be unexpected. Especially when two objects touch with strong forces, the simulation can overreact, which leads for example to the robot falling or objects flying through the air. These failures can be avoided to some extent by careful control of the robot and the choice of good positions for the manipulation actions. However, the subjects were not instructed explicitly on these details.

2.2.2 Scenarios

All tasks were given to the subjects in visual and written textual form as presented in Figures 6–8. In simulation and in reality, the objects were coloured (in simulation they were covered completely by one colour, in reality white dishes and metal cutlery were furnished with coloured markers as shown in Figure 5).

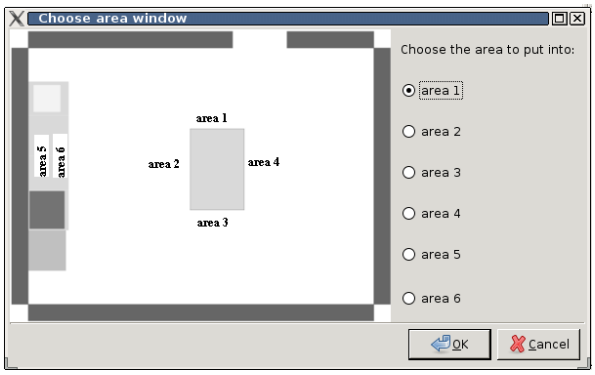


Figure 4: Predefined areas in one of the available kitchens.

Figure 5: Marked objects in the real-world trials.

In simulation, the subjects were given four simple tasks that were performed as a first exercise in the study. In two of them, the subject had to move a plate, in the others a cup was to be manipulated. The simple tasks were all performed in Kitchen 2.

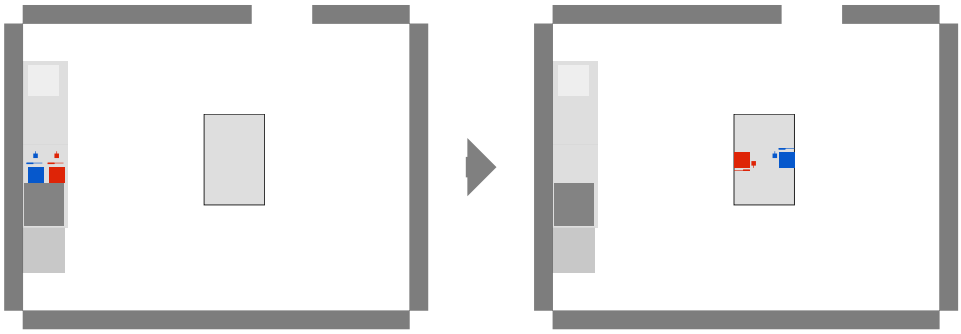
The complex tasks that were performed both in reality and simulation are numbered ct-1-1 through ct-1-6. Three tasks were to set the table (Figure 6) and three to clear the table (Figure 7). There were two tasks in which the subjects had some freedom in the choice of objects and their positions. In scenario ct-1-6 the final positions of the objects could be chosen by the subjects, the objects to be used were given by the task itself. Setting the table in scenario ct-1-3 allowed the choice of two complete place settings out of three available place settings and the goal positions were only specified without colours.

In simulation, there were two additional scenarios, which are presented in Figure 8: one is setting the table, the other clearing it. In both tasks the object positions were completely specified.

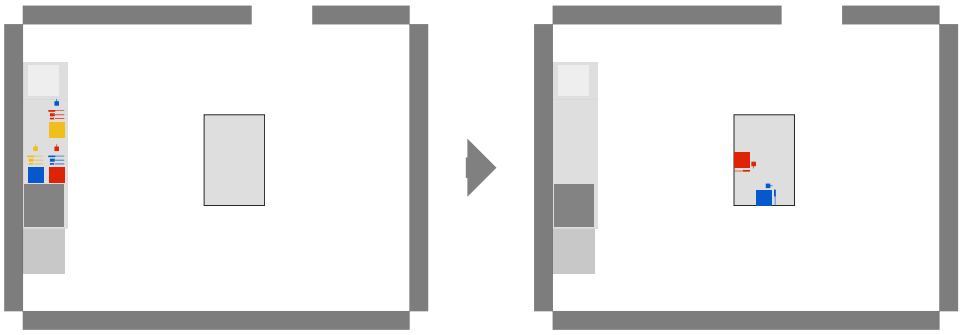
In total, we collected data of 54 trials in the real world and 108 in simulation (36 simple tasks, 72 complex tasks). Because of unrecoverable failures, 8 trials of complex tasks had to be repeated. We only consider the successful trials in our evaluation.

2.3 Data Analysis

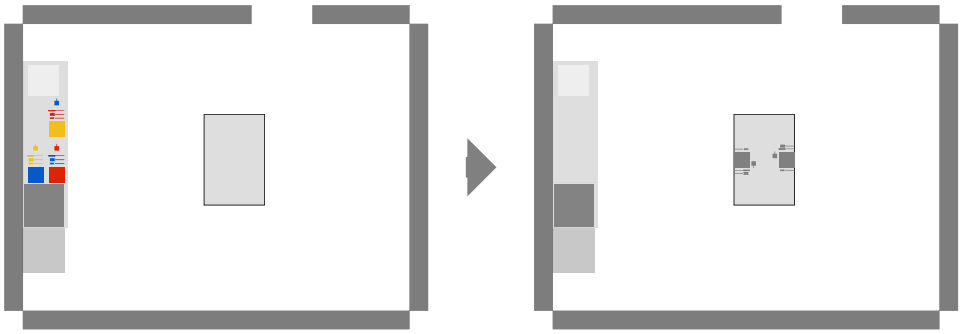
For all trials in simulation and reality we analysed data of complete tasks and particular subtasks. Moreover, we evaluated data over all trials, which includes user feedback.



(a) Task ct-1-1: place red breakfast cover (plate, cup, knife) at left long side of the table and blue breakfast cover at the other long side of the table.

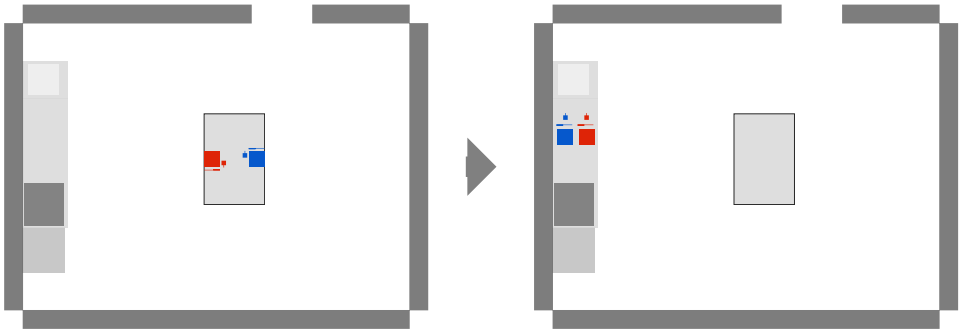


(b) Task ct-1-2: place red breakfast cover (plate, cup, knife) at left long side of the table and blue breakfast cover at bottom short side of the table.

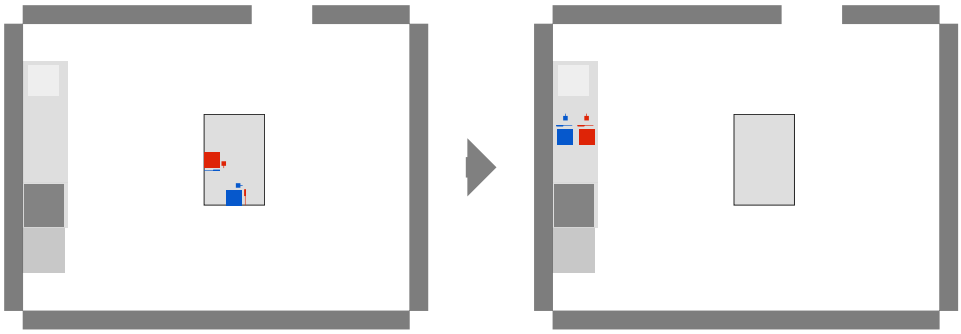


(c) Task ct-1-3: place two complete covers (plate, cup, knife, fork, spoon) on the long sides of the table.

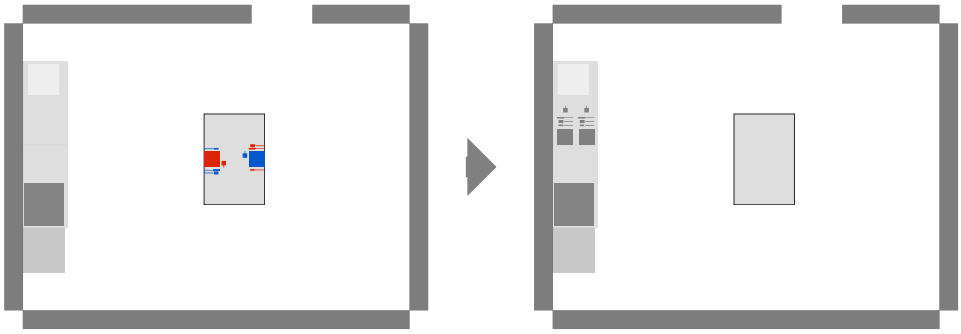
Figure 6: Table setting tasks to be performed in real and simulated kitchen.



(a) Task ct-1-4: put items from table next to the sink, back row: blue items, front row: red items.

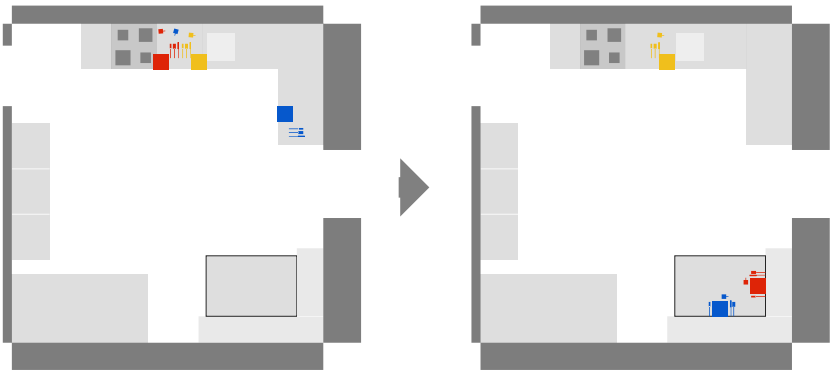


(b) Task ct-1-5: put items from table next to the sink, back row: blue items, front row: red items.

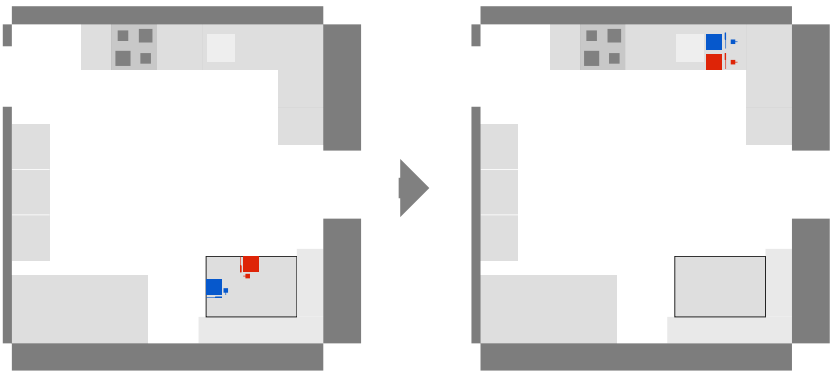


(c) Task ct-1-6: put items from table next to the sink.

Figure 7: Tasks for clearing the table to be performed in real and simulated kitchen.



(a) Task ct-2-1: place complete red cover (plate, cup, knife, fork, spoon) at right short side of the table and complete blue cover at the bottom long side of the table



(b) Task ct-2-2: put items from table next to the sink, back row: blue items, front row: red items.

Figure 8: Complex tasks to be performed in second simulated kitchen.

2.3.1 Data of Basic and Compound Actions

We considered some data on the level of actions. This includes the basic actions *grip* and *put-down* provided as an action to the user. Besides, we defined a *carry task* as taking two objects, carrying them, and putting them down. In simulation, a carry task is characterised by the following data:

- object carried in left hand: object type $otype_l$, object colour $ocol_l$;
- object carried in right hand: object type $otype_r$, object colour $ocol_r$;
- hand which was used for the first grasp;
- hand which was used for the first put-down action.

Almost all actions in the simulated trials can be assigned to a carry task. One exception is a trial in which the subject only used one hand for fulfilling a task, showing only "semi-carry tasks". This trial was not used in the evaluation of carry tasks. Another behaviour shown by one subject also fell out of the carry task schema: presumably to avoid failures, some objects were not carried to their goal positions directly, but were put at an intermediate position and moved later for finishing the complete task. In this case, the first manipulation of the object was counted in the respective carry task and the second movement was not used in the analysis.

For the tasks performed in reality, we only identified carry tasks in those trials where the subjects were restricted to take only one object per hand. Because in reality, people grip and put down objects at the same time, the carry tasks are only defined by

- object carried in left hand: object type $otype_l$, object colour $ocol_l$; and
- object carried in right hand: object type $otype_r$, object colour $ocol_r$.

Object Preferences We identified specific preference types for the objects used in a carry task:

1. Object preference: Carrying objects of the same or similar type together.
   (a) Strong object preference: Taking objects of the same type in a single carry task: $typeeq \Leftrightarrow otype_l = otype_r$
   (b) Weak object preference: Taking objects of similar type in one carry task: $typesim \Leftrightarrow \exists c.typeclass(c) \wedge member(otype_l, c) \wedge member(otype_r, c)$, where the type class c can be either of the two sets $dishes = \{cup, plate\}$ or $cutlery = \{knife, spoon, fork\}$
2. Colour preference: Taking objects of the same colour in a single carry task: $coleq \Leftrightarrow ocol_l = ocol_r$

The predicates *typeeq*, *typesim* and *coleq* were determined for each carry task, as well as the compound preferences $typeeq \vee typesim$ (some object preference) and $typeeq \vee typesim \vee coleq$ (colour or object preference).

Failures Except for the occasional disregard of the restrictions asked of the participants in the real-world trials, failures only occurred in the simulated trials. For the analysis of failures in the simulation, we also considered single grip and put-down actions as well as complete carry tasks. For all three kinds of actions, we counted the number of tasks n_p^m of this kind (m can take the values $\mathcal{G}$ for grip, $\mathcal{P}$ for put down and $\mathcal{C}$ for carry) observed over all trials of one participant p. Then we counted the number of failed tasks f_p^m and calculated the percentage of failed actions for each action type and participant:

$$F_p^m = \frac{f_p^m}{n_p^m}$$

The average value for each action type is denoted as $\bar{F}^m$, the standard deviation as $\hat{F}^m$.

Gripping Parameters In the simulation data, we evaluated some parameters of the grip action. One parameter is the distance $d_{o,p}$ between the hand-controlled robot and the object to be gripped. o is the object type (plate, cup, knife, spoon or fork) and p the participant. From this data, we calculated the average distance $\bar{d}$ over all objects and participants as well as the standard deviation $\hat{d}$.

Besides, we calculated the absolute value of the rotation angle $\phi_{o,p}$ of the simulated robot relative to the line of sight between the robot and the object. Again we used the average $\bar{\phi}$ and standard deviation $\hat{\phi}$.

2.3.2 Task Data

We used quantitative and qualitative measures for complete tasks, i.e. the execution of one scenario until its goal configuration was reached. This analysis of the behaviour in reality and simulation included the duration needed for completing the tasks and the objects that were chosen for the partially defined goal configurations of Scenarios ct-1-3 and ct-1-6. Besides, for real-world execution, we considered the unconscious violations of the restriction to carry only one object per hand. Finally, we classified the paths on which the subjects navigated the robot in simulation.

Duration of Tasks For analysing and comparing the durations of tasks, we measured the times $t_{s,p}^m$ for each scenario s performed by each subject p. The parameter m can take the values $\mathcal{S}c$ for complex tasks in simulation, $\mathcal{S}s$ for simple tasks in simulation, $\mathcal{R}n$ for all scenarios performed in reality, and $\mathcal{R}r$ for those trials in reality where the subjects handled only one object per hand. For each scenario s, we calculated the average value $\bar{t}_s^m$ and standard deviation $\hat{t}_s^m$ both for reality and the real world.

From these values, we calculated several parameters for different evaluations:

- the time needed by each participant p normalised by the average for a certain scenario s: $T_{s,p}^m = t_{s,p}^m / \bar{t}_s^m$;
- the average of the weighted time values for each participant p:

$$\bar{T}_p^{\mathcal{R}n} = \left(\sum_{s=ct-1-1}^{ct-1-6} T_{s,p}^{\mathcal{R}n} \right) /6 \ , \ \bar{T}_p^{\mathcal{R}r} = \left(\sum_{s=ct-1-1}^{ct-1-6} T_{s,p}^{\mathcal{R}r} \right) /4 \ ,$$

$$\bar{T}_p^{\mathcal{S}c} = \left(\sum_{s=ct-1-1}^{ct-2-2} T_{s,p}^{\mathcal{S}c} \right) /8 \ , \ \bar{T}_p^{\mathcal{S}s} = \left(\sum_{s=st-1}^{st-4} T_{s,p}^{\mathcal{S}s} \right) /4$$

- the weighted time of each participant for each scenario normalised by the participant's average performance $\tau_{s,p}^m = T_{s,p}^m / \bar{T}_p^m$
- the standard deviation needed for each scenario normalised by the average time of this scenario: $\hat{T}_s^m = \hat{t}_s^m / \bar{t}_s^m$

Object Choices For the two scenarios in which the goal positions of the objects were not completely specified (ct-1-3 and ct-1-6), the goal configurations were classified along two criteria:

1. Order strategy: Move a place setting as laid out in the original position. For example, in scenario ct-1-3 a popular combination was a place setting made from red plate, blue cutlery and red cup together with a place setting composed of yellow plate, red cutlery and blue cup (which corresponds to the first row of objects in the original configuration).
2. Sorting strategy: Sort objects according to their colour.
   (a) Strong sorting strategy: Put only dishes of the same colour in one place setting.
   (b) Weak sorting strategy: Use at most two colours for setting two place settings (only relevant in ct-1-3).

Restriction Violations In the restricted trials (carrying only one object per hand) executed in reality, we evaluated if this restriction was violated. Because of the low number of trials (36 restricted trials over all subjects) and the different severity of violating these constraints (carrying more than two objects at a time is further away from the instructions than taking only two objects, but changing the hand for an object), this data was only used as evidence for qualitative evaluation of the need for additional actions for the human-controlled robot.

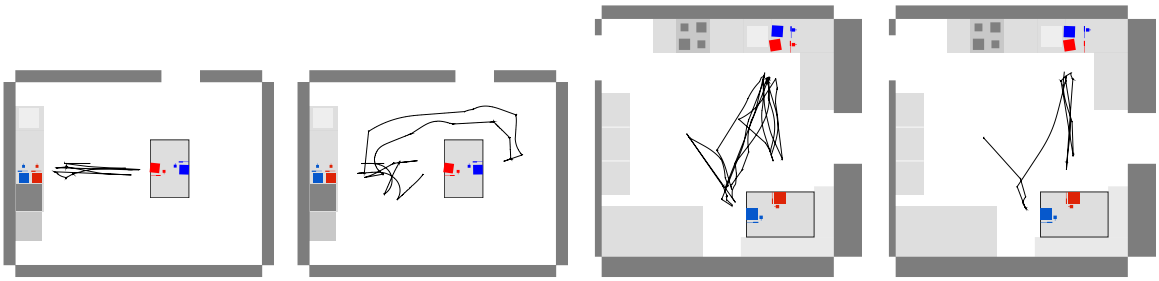
Navigation Paths We analysed the paths that the subjects used in both kitchens qualitatively along three categories:

C1. the subject takes the direct way;

C2. the subject accepts a long way (in Kitchen 1 around the table, in Kitchen 2 next to the narrow side of the table);

C3. no clear assignment to category 1 or 2 possible: for example when the subject moves a slightly longer way than the direct path, but not all the way around the table.

Figure 9 shows some examples for the classifications. The paths in scenario ct-2-1 were not classified, because they were very similar for all participants and would all have fallen into category 3.



(a) C1: direct (b) C2: long way (c) C2: long way (d) C3: mixed

Figure 9: Qualitative classification of paths — examples for both kitchen environments.

2.3.3 User Data over all Trials

Apart from aggregated data of single trials or carry tasks, we obtained the evaluation of users concerning their own experiences over all tasks in simulation and reality. The data consisted of statements to be evaluated on a five-level Likert scale (see Figure 12) and four open questions:

1. What was difficult in using the simulation control?
2. Which actions, except the available ones grip and put down, should the robot in the simulation be able to perform?
3. In your opinion, which differences are there between the execution of the tasks in the simulator compared to the real world? What did you do differently?
4. Which strategy did you have for setting and clearing the table? Was the one in the simulation different from the one in the real-world execution?

3 Results

We first present the results for the usability of our simulation environment. These results are highly dependent on the underlying simulation software and its control options. They are the foundations for the more abstract results on the comparison of how people perform household tasks in reality and simulation, where we compare the high-level behaviour in scenarios ct-1-1 to ct-1-6. These results are more general and could probably be reproduced using other simulation environments. Third, we provide a brief comparison of the tasks performed in simulation in the two different kitchens.

3.1 Usability

The first part of the evaluation considers the usability of the used simulation control. In this section, we use the data of all scenarios performed in simulation.

3.1.1 Agility in using the simulation control

We evaluated the skillfulness of the subjects in handling the simulation along two dimensions: speed and failures.

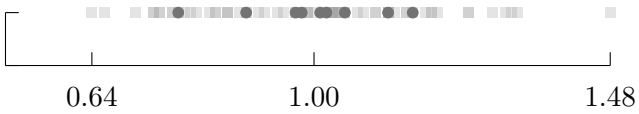


Figure 10: Relative times $T^{\mathcal{S}}_{s,p}$ needed by users for the trials compared to the average time subjects needed to fulfil the same scenario (light grey squares) and the average per participant $\bar{T}^{\mathcal{S}}_{p}$ (dark dots).

Speed The speed with which users achieved the tasks can only be compared between the users. Ideally, there should be only a small difference in the times needed to fulfil a task for each user. Figure 10 shows the normalised times $T^{\mathcal{S}}_{s,p}$ indicating the deviation of each participant p from the average performance for each scenario s (square marks). The maximum deviation is 48%, the variance of the whole data set is 18%. Shown as darker dots in Figure 10 are the average deviation values $\bar{T}^{\mathcal{S}}_{p}$ for each subject. The maximum average score is 16% slower than average, the fastest subject was 22% faster than average.

Even though a variance of around 20% seems a lot, we compared the variance in simulation to that of table setting and clearing in the real world for the scenarios that were performed in both worlds (see Section 3.2.2) and found that the deviation in reality amounts to the same relative variance. This means that the differences in speed can be attributed to the variance in thime that is normal for t he kinds of tasks we observed. Therefore, the usability of controlling the robot seems satisfactory with respect to speed.

Failures For the evaluation of failures in the execution we have to consider the underlying software. The realistic simulation of physical processes is still subject to research. Therefore, it is not surprising that the ODE library we are using sometimes produces events that would not occur in the real world. In particular, when the robot hits the table with its arm, the simulation "overreacts" and the robot can fall down and lose the object, even after seemingly small impacts. These kinds of failures can only partially be avoided by a careful control of the robot and a good estimation of its capabilities.

Besides, the implementation of the robot's grip and put-down actions is not based on sophisticated path planning algorithms, but makes heavy use of heuristics. This works well most of the time, but can cause failures that are not expected by the user. And the robot doesn't take care at all to avoid collisions with its arms.

With these words of caution in mind, Table 2 gives an overview of the failures for gripping and putting down objects as well as for complete carry tasks. Activating the gripping behaviour failed on average in about 4% of all gripping tasks. The reason for a failed gripping task is most of the time that the robot is in a position from where it cannot grip the object (i.e. the automatic gripping routine doesn't find a path), usually it's too far away. It can also happen that the robot crashes into a piece of furniture and falls down before it can grip the object.

With an average of 9%, the failure rate for put down tasks is more than twice as high as for gripping. The causes for failing are similar to those when gripping with the additional problem that the object positions near the wall are very hard to reach and the robot sometimes hits the wall with the objects.

Even though the failure rates for single grip and put-down actions are not too high, they sum up when considering complete carry tasks. Taken over all subjects about one in five

Table 2: Percentage of failed actions. The columns min and max show the minimum and maximum failure rates of the individual subjects.

	min	max	average	standard deviation
grip	0.0%	9.4%	4.18%	3.31%
put down	4.4%	15.2%	9.46%	3.49%
carry	9.7%	28.1%	20.97%	5.45%

carry tasks included at least one failure. Sometimes these failures could easily be recovered by repositioning the robot and retrying the action. This was not possible when the robot had lost objects in the attempt to put them down. Considering the subjects' own impression of their behavior (Section 3.1.3) and the object-related observations (Section 3.2.1) the failures didn't have a strong effect on the overall behavior and planning schemes of the subjects.

Overall, the failure rate is acceptable for a study investigating high-level behaviour like the present one, but shows potential for improvement. We expect the failure rate to be lower when users are instructed explicitly on the physical abilities of the robot. And after the findings of the user study we have implemented a warning mechanism that tells the user before an action that the robot might hit the furniture.

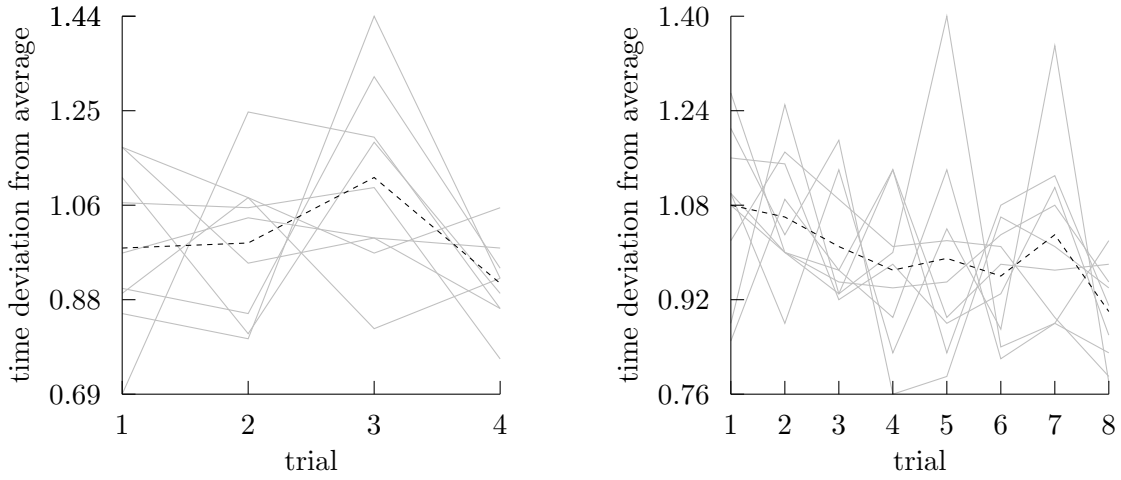
3.1.2 Improvement

For using the simulation in further experiments, we wanted to know if the ability to use the simulation increases over time and how much training people need until they can handle the simulation sufficiently well. The first four tasks to be performed in simulation contain only one object and the instruction to bring it to another predefined place. Since it only contains one carry task, these trials are less noisy with respect to failures. Besides, they are the first trials for all subjects and can give an indication on how long it takes until subjects converge to their full capabilities with the simulation.

Figure 11a shows the relative trial durations weighted with each participant's own average $\tau_{s,p}^{\mathcal{S}s}$ for all participants. The scenario s differs for each trial depending on the participant, because the trials were performed in random order. Time deviation values above 1 indicate that a task was performed slower than the subject's average, values below one are faster than the personal average. The dashed line shows the average over all participants for each trial number (which corresponds to different scenarios for different subjects).

For all participants, the performance *decreases* at some trial. This can be explained by the fact that the subjects had very little experience in handling the simulation (5 minutes try-out time for each participant before the start of the experiment and no instruction indicating good grasping positions and other parameters). It seems that when people felt sure enough with the basic functionality, they started to try new things. One instance for this is the use of the functionality to stop only the robot's rotation and keep its translational velocity. Several subjects didn't use this functionality in the first run, but started to try it later. The fourth trial has a value below their average performance for almost all subjects. This might indicate that after four trials most people have adapted well to the simulation.

Overall, these results for the development of skills in the simple tasks should be taken with a grain of salt. The scenarios were very short and small detours in controlling the robot could have huge effects in these statistics. And we don't know how the development would



(a) First four trials (simple tasks).

(b) Complex tasks.

Figure 11: Time needed for each participant p_i compared to average performance of this participant p_i per trial. Note that the trial numbers don't correspond to the same scenario for each participant, since the trials were chosen randomly. This means that scenario st-1 can be the first trial for one participant and the third for another. The dashed line indicates the overall trend (average value for all subjects per trial number).

have been in further trials of this kind of tasks.

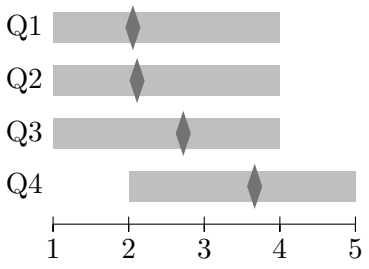
Figure 11b shows the same type of diagram for the eight complex tasks $\tau_{s,p}^{\mathcal{S}c}$. In these scenarios, deviation from the average can have many reasons apart from the user's ability to handle the simulation, like failures caused by errors in the physical simulation or distractedness of the participant. Still, the results indicate a stable ability to handle the simulation with a slight improvement over time.

Even though the measurements are noisy, we conclude from these observations that people should perform several pick and place tasks to get accustomed to the control. It seems helpful to provide subjects with some insight into the robot's capabilities and limitations. For this experiment, we only explained the pure workings of the simulation, but let the users find out where they have to position the robot in order to grasp an object or put it down. Especially the fact that the robot's arms are extendable is not obvious and the learning process for users could probably be accelerated by telling them these facts.

3.1.3 User Self-Evaluation and Satisfaction

After performing the tasks in reality and simulation, we asked the subjects about their experience with the simulation. Figure 12 shows the range and average value of answers obtained. Besides, we asked what the subjects considered as most difficult in handling the simulation.

Most participants felt that they can handle the simulation well and achieve the tasks quickly. Only two subjects didn't agree (scores higher or equal 3) that they handled the simulation well (Question 1) and three disagreed to having achieved the tasks quickly (Question 2). The data didn't show any significant relationship between the recorded data and the self-evaluation of the subjects. However, the subjects who rated their abilities very low were in fact among the slower participants.



Q1) I feel that I can handle the simulator well.

Q2) I think that I was able to achieve the tasks in the simulator quickly (as quickly as the simulated robot allows).

Q3) The simulated robot executes the actions (grip and put down) without errors.

Q4) The simulated robot executes the actions fast.

Figure 12: Answers to questions about the subjects' own evaluation of their skills (questions 1 and 2) and their assessment of the simulated robot's skills (questions 3 and 4). The Likert values were labeled: 1=fully agree, 2=partially agree, 3=don't know, 4=partially disagree, 5=fully disagree.

When asked about the difficulties in handling the simulation, a common answer was the difficulty in estimating the robot's capabilities, especially its gripping radius. Some people explicitly mentioned that it would have helped them to get more explanation before the experiment.

Two participants had problems steering the direction from the robot's point of view. Although there is a window with the robot's view of the world, everyone concentrated on the bigger window where the whole scene was shown from an outside view (cp. Figure 3). But the measurements didn't show any significant disadvantages for these subjects, because they often reconsidered their actions quickly and were able to control the robot satisfactorily. It would be interesting to investigate if providing a bigger window of the robot's own view can help people who have difficulty with the geometry.

The results on the question about the robot performing tasks reliably was surprisingly good. Considering the that on average every fifth carrying task contained at least one failure (of varying severity), four subjects fully or partially agreed that the robot performs actions without errors and there was no complete disagreement to the statement. This indicates that the failures in the pick and place tasks didn't have a strong effect on the overall strategies of the subjects. The average score of 2.7 is in line with the measured values for failures — not disastrous, but with potential for improvement.

The question about the robot's speed was rated clearly between "don't know" and "fully disagree". Only one participant could "partially agree". This result is not surprising when considering that table setting in the simulation takes about 20 times as long as in reality. The most boring parts for the participants are the grasping and putting down actions, where the user can only wait until the robot has finished.

To sum up, the usability of the used simulation was sufficient for abstract behaviour observation in the present study. For more focused studies, the failure rate could be problematic and should be minimised by better instructions and more sophisticated action execution. Besides, the slowness of the simulation must be taken into account when designing an experiment.

3.2 Comparison of Task Execution in Simulation and Reality

For the comparison of the results in the real world and in simulation, we only use scenarios ct-1-1 to ct-1-6. And we only use those trials of the real-world data where people were restricted to carry one object per hand.

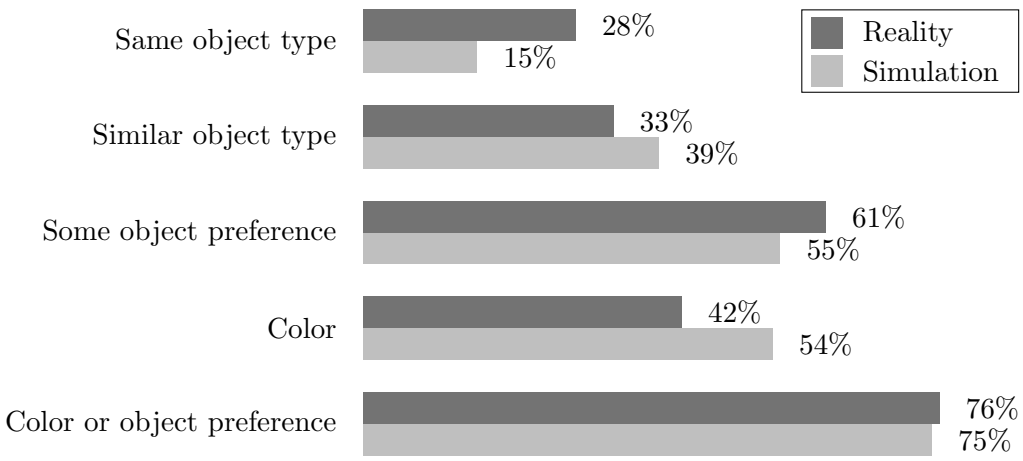


Figure 13: Comparison of preference types over all subjects.

3.2.1 Preferences in Object Handling

Object Preferences in Carry Tasks Our subjects showed certain preferences for the objects they take in a single carry task. We examined the preferences defined on page 10 and compared their applicability in the real-world task execution to the one in simulation.

Figure 13 shows the proportion of carry tasks that can be explained by each of the preferences, using the data of all subjects and scenarios ct-1-1 through ct-1-6.

It shows that preferences in carry tasks vary only little between execution in reality and in simulation. In the real world, people prefer to carry identical objects, whereas in simulation objects of the same type are more often carried at the same time. But in both cases carrying objects of the same or similar kind can explain about half of the observed carry tasks, being slightly stronger in real-world execution. An explanation might be that in real-life execution it is easier to grasp objects of the same kind at the same time than taking two objects with different grasps. In simulation, this motorical preference doesn't play a role for efficiency. However, it seems that people still like to carry objects of the same kind (one subject even complained in scenario ct-2-1 that it was not possible without loss of efficiency to stick to the weak object preference).

Similarly, taking objects of the same colour can explain about half of the carry tasks, in the real world slightly less. Taken together about three in four carry tasks can be explained by object or colour preferences — both in reality and in simulation. Taking into account that the control of the simulated robot is a lot slower than acting in the real world, it is quite surprising that the preferences in simulation account for approximately the same percentage as in real-world execution. The preference for taking objects of the same or similar kind that are sacrificed in simulation for efficiency reasons seems to be compensated by the urge to take objects of the same colour.

Another indication that the preference types we defined matter in the task execution of humans are the observations from the real-world trials in which the subjects were allowed to use all their abilities to fulfill the task. There were three principle strategies:

1. carrying all objects at once, either by arranging them in one stack or by building a partial stack of plates and cups and carrying the cutlery together with the stack;

2. carrying all objects at once using both hands separately;

3. going the way between the worktop and the table several times.

In strategies 2 and 3, we can observe object and color preferences in addition to the joint carrying of objects belonging to one place setting. When using the second strategy, in one trial the subject carried the a stack of all red objects in one hand and a stack of all blue objects in the other. In two cases, the plates and knives were carried in one hand, the cups in the other, indicating a type preference.

When using strategy 3 a common pattern was to carry all objects of one color first and the remaining objects in the second move. This behavior was an instance of carrying the objects of one place setting at a time. No subject ordered the objects by color before carrying them. In one trial for scenario ct-1-3 the subject carried the plate and cutlery of the first place setting first, in the second move both cups were transferred and in the last move, the second place setting was completed. This behavior also indicates a preference of types as well as the completion of place settings.

Of course, the observed preferences are no surprise. But for a robot it is interesting to learn such human preferences. And if these are similarly strong in simulation and reality, the simulation turns out to be an appropriate testbed for our research on human-robot collaboration.

Strategies for Object Choices Similarly to preferences of carry tasks, we examined which objects were chosen and at which places they were put when the goal position was not fully specified, i.e. in scenarios ct-1-3 and ct-1-6. In scenario ct-1-3 the subjects could choose two complete place settings out of three complete place settings to put on the table, in ct-1-6 the two place settings on the table were to be put at places to be chosen next to the sink. For classifying the choice of objects we use the strategies "order" and "sorting" explained on page 12.

Again, both kinds of strategies are found in the real world and in simulation. When confronted with material for three place settings and only setting the table for two people (scenario ct-1-3), the impulse for setting place settings of one colour is stronger in the real world. In simulation, only the weak sorting strategy could be observed. One reason might be that people wanted to change their behaviour on purpose in order not to repeat what they had done in the real world, possibly with the additional condition that setting the table in simulation is more boring and leaves more time to think about such things. Another explanation, which was also mentioned by the subjects themselves, could be the intricacies in handling the simulation and the slowness of the manually controlled robot. In simulation, people tried to speed things up as much as they could.

Although most subjects showed an observable strategy, only a minority employed the same strategy in both worlds. Possibly, both strategies are regarded as similarly natural and people don't care which of them they use. However, this point can only be clarified in further experiments with repeated trials.

Another interesting detail is that in the simulated trials of scenario ct-1-3 three subjects chose exactly the same goal configuration using mixed place settings of red and yellow dishes and cutlery. Only one subject produced the same result configuration (with only changing the sides of the table) for simulation and real world in this scenario.

Table 3: Comparison of strategies in real world and simulation for choosing objects and their positions in scenarios ct-1-3 and ct-1-6. The table contains the number of subjects choosing each of the strategies. The bottom part shows how many subjects used the same strategy in simulation and the real world without necessarily using the same objects ("Keep strategy") and those who employ different strategies in both worlds. The "Keep strategy" criterion also holds when a person changes from strong to weak sorting strategy. Note that in some cases the subject didn't follow any of those strategies.

	ct-1-3		ct-1-6	
	real	simulation	real	simulation
Order	3	3	5	6
Strong sorting	5	0	4	3
Weak sorting	0	4		
Some strategy	8	7	9	9
Keep strategy	2		4	
Change strategy	5		5	

3.2.2 Comparison of Time Scales

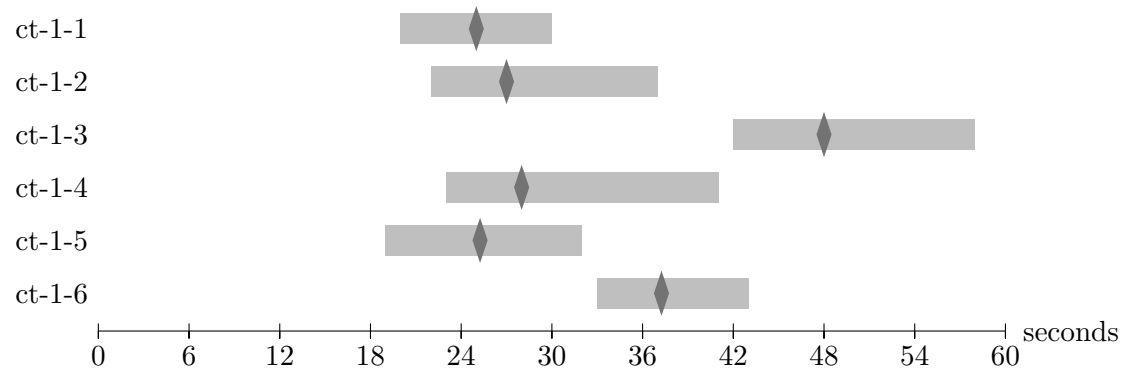
To get an intuition of the different time scales in reality and simulation we compared the durations and their variances needed for completing table setting or clearing tasks. Figures 14a and 14b show for each task ct-1-1 through ct-1-6 the range of normalised times $T_{s,p}^{\mathcal{S}c1}$ and marks the average of these values per scenario[1]. Table 4 shows some quantitative measures for comparing the deviation of the candidates.

Our data shows that even with errors occurring in the simulation and different skills in navigating the robot, the overall results are very similar to those observed in the real world. Qualitatively looking at Figures 14a and 14b shows that scenarios ct-1-1, ct-1-2, ct-1-4 and ct-1-5 are approximately on one time scale (which is not surprising as they involve the same number of objects) and scenarios ct-1-3 and ct-1-6 (involving more objects than the other scenarios) share another time scale and larger variances.

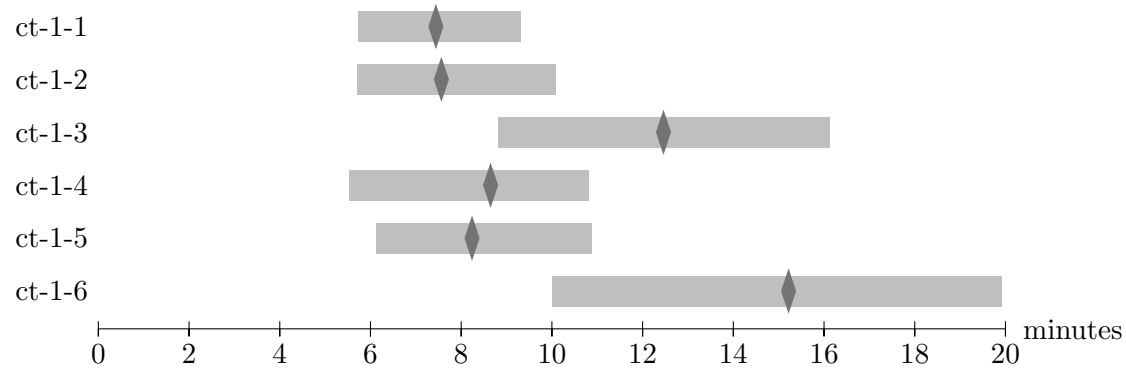
For the scenarios ct-1-3 and ct-1-6, which involve more objects, there is an obvious difference in the timing. Whereas in reality ct-1-3 took a lot longer to complete than ct-1-6, the reverse is true for simulation. Since in both scenarios the same number of objects had to be moved, it is surprising that there were any visible differences at all. It is possible that in the real-world trials scenario ct-1-3 was more complex, because the objects to be moved had to be chosen by the subjects, whereas in ct-1-6 only the goal positions had to be decided on. Any such complexity of object choice in ct-1-3 would disappear in simulation, because the subjects had lots of time to choose objects compared to the slowness of executing the actions. The longer duration of ct-1-6 in simulation might be due to a higher failure rate, because the goal positions of the objects near the wall were generally harder to reach for the robot than goal positions on the table. This would also explain the high variance in completion times for ct-1-6 in simulation.

Table 5a shows the average times $\bar{t}_s^{\mathcal{R}r}$ and $\bar{t}_s^{\mathcal{S}c1}$ and the standard deviations $\hat{t}_s^{\mathcal{R}r}$ and $\hat{t}_s^{\mathcal{S}c1}$

[1]The parameter indicating the world $\mathcal{S}c1$ denotes the complex tasks performed in Kitchen 1 (i.e. scenarios ct-1-1 – ct-1-6). Likewise, $\mathcal{S}c2$ means scenarios ct-2-1 and ct-2-2 in Kitchen 2.



(a) Measurements in reality (using only data of those trials where only one object per hand was to be carried).



(b) Simulation results.

Figure 14: Times for completing tasks ct-1-1 through ct-1-6 in the real world and simulation showing the range of times (bar) and the average time (diamond-shaped mark).

Table 4: Comparison of time deviation for completing tasks in reality and simulation. For the real world the times are used only from those trials where the person was restricted to using the hands like a robot.

	real world (seconds)	simulation (minutes)
average time	31.78	9.93
standard deviation	10.2	3.51
relative deviation	32.1%	35.4%

(a) Comparison of average time $\bar{t}_s^m$, standard deviation $\hat{t}_s^m$ and relative deviation $\hat{t}_s^m/\bar{t}_s^m$ for $m = \mathcal{R}r$ and $m = \mathcal{S}c1$.

	ct-1-1	ct-1-2	ct-1-3	ct-1-4	ct-1-5	ct-1-6
real	17.6%	18.8%	13.4%	24.7%	16.6%	13.6%
simulation	13.7%	21.1%	19.9%	18.0%	18.5%	21.1%

(b) Comparison of relative deviation $\hat{T}_s^{\mathcal{R}r}$ and $\hat{T}_s^{\mathcal{S}c1}$ for each scenario.

over all scenarios and subjects. Those measures are only very rough numbers, because the scenarios vary in the number of objects. Still, the relative deviation shows that in both cases the relative variance is around 30–35%. The more detailed numbers in Table 5b show a similar picture. All relative variations are around 10-20%, both in reality and in simulation.

These results show that — taking into account a large scaling factor — the relative times of activities are comparable between real-world behaviour and simulation. However, the durations of the single actions cannot be expected to scale in this manner. This detail could only be clarified in a more focused study where the start and end of actions is clearly defined.

3.2.3 User Experience

Beside the quantitative measurements, we also wanted to know how the subjective feelings of the participants were towards the differences of acting in the real and simulated world. In this section, we consider the answers given to questions 2, 3 and 4 shown in Section 2.3.3 on page 13. In addition, we asked to evaluate the remark "Because the simulation doesn't allow stacking of objects, the whole activity becomes very unnatural." on a five-level Likert scale.

Indeed, people missed the ability to stack objects. The average Likert score was 2.2 (2 corresponds to "partially agree"). In the open question about desired additional actions, four people named stacking and three mentioned this difference in question 3 (which might however be influenced by the explicit question about stacking). In the same direction, three subjects would like the robot to grip two objects at once (which is indeed what all subjects did in the real world) and some subjects mentioned during the trials that it would be helpful if the robot could hold more than one object in a hand and would be able to hand over objects from one hand to the nother.

Comparing these opinions to the execution of the tasks in the real kitchen is somewhat surprising. Some people complained about not being able to stack objects in simulation, although they didn't stack the objects when performing the tasks in reality (the two trials in which they could choose their actions freely). And no one tried to stack objects when they were asked not to do so, which indicates that stacking is a conscious activity that people can live without easily. In contrast, the urge to grasp several objects at a time or handing them over was violated more or less heavily nine times in 36 trials and the subjects mostly weren't aware of the rule violation.

When asked about the perceived differences of reality and simulation and the different strategies, the most common answer was that it was a lot harder to estimate the physical constraints and the robot's actions in the simulation. Besides, a common impression was that people stray from their preferences of taking objects of the same type or colour when working in simulation.

The first point can definitely be confirmed by the observations. As described above, the physics engine ODE can overreact to collisions between objects, which can quickly lead to the robot falling down or losing objects. Besides, even though the robot's arms are designed for simulation and are more agile than most real currently available arms, it is still a lot harder to evaluate good positions for a robot to stand compared to human abilities. Although this provides a nice showcase for educational purposes (i.e. demonstrating non-scientists how difficult the control of autonomous robots is), it makes the control of the robot difficult.

The second comment on the preferences includes self-observations such as "In real execution I always take things which are of the same colour. But for the robot I always try to do it as easily as possible." or "I normally put the biggest objects first: plates, cups and

so on. I tried to do the same in the simulator, but in the simulator (to save some time) I tried to be aware of the hand with which I was grasping the objects and which would be its last position." Interestingly, the results in Figure 13 suggest that people do not change those habits significantly, even though the control of the robot in simulation is more difficult. However, Figure 13 also shows that people tend to grasp objects of the same type in the real world, whereas in simulation, similar objects are more often carried together. And the colour preference is even more pronounced in simulation than in real-world execution. Also the strategies of which objects to put on the table as shown in Table 3 suggest that there is some more emphasis on efficiency in the simulation, but that similar strategies (in this case weak sorting strategy as opposed to strong sorting strategy) are employed in simulation as well.

One explication for this discrepancy between felt and observed differences might be that people would like to take some constellation of objects, but "replan" their activity when they realise that this might not be efficient in the simulator. And then the new course of action does show some preferences, but not the ones that the subjects had originally intended. Furthermore, in the simulated world people have much more time to think about their actions. Whereas in real life, activities are performed unconsciously, there might be an attempt to find an "ideal" way in the simulation.

In all, the behaviour of people setting and clearing the table is comparable in simulation and reality with respect to the object choices and the time scale of the whole task. The main difference seems to be the more conscious execution of the tasks in simulation, which seems to be due to the slowness of the simulator.

3.3 Comparison of Behaviour in two Simulated Worlds

One of the advantages of using simulation for human-robot experiments is the possibility to create different virtual worlds for the user to act in. Our experimental scenarios ct-2-1 and ct-2-2 are in a different kitchen than the ones in ct-1-1 through ct-1-6. In the following we compare the results in the two kitchens. Given the scarcity of the data (only two scenarios in Kitchen 2), these results can only be indicative.

3.3.1 Carry preferences

Parallely to the comparison between reality and simulation, we considered the preferences for similar objects in carry tasks for the two kitchens. Figure 15 shows the ratios of tasks, where a specific preference was shown.

The general picture of users having preferences to carry objects of the same or similar type or of the same colour can also be observed in the second kitchen. But the differences in preferences are higher between the two simulated environments than between reality and simulation in Kitchen 1 (cp. Figure 13). Especially the preference for carrying objects of the same colour, is a lot more pronounced in Kitchen 2.

This might be explainable with the specific configuration of objects, especially in scenario ct-2-1, where most of the blue objects are grouped at one place. It would be interesting to investigate this phenomenon more closely in a more focused study.

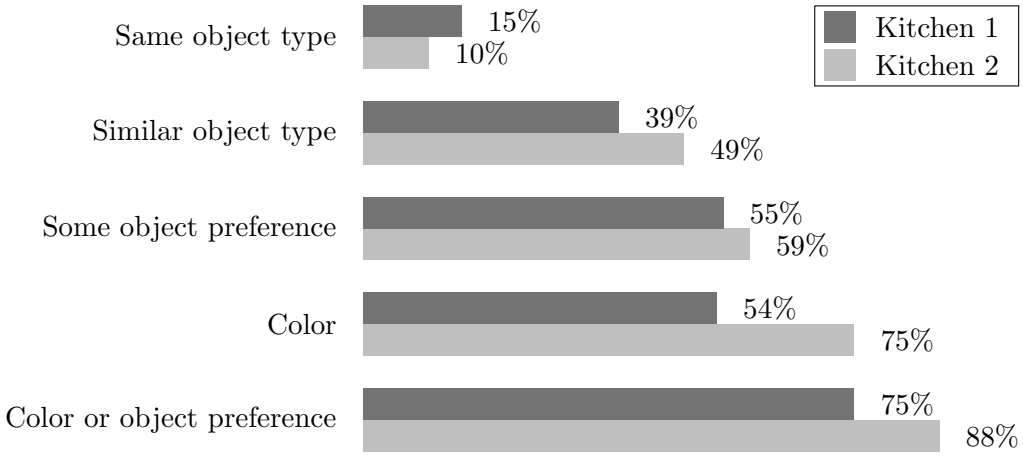


Figure 15: Comparison of preference types over all subjects in the simulated kitchens.

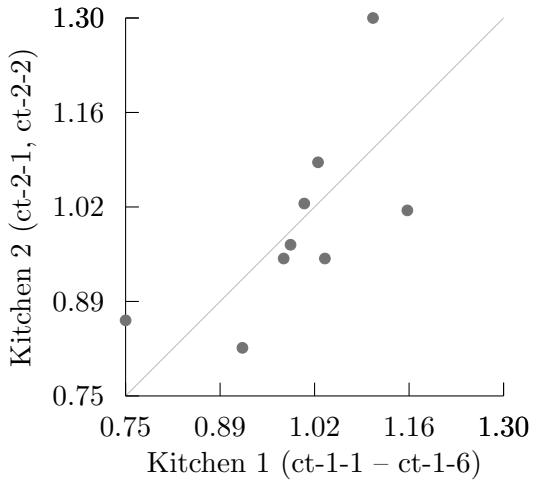


Figure 16: Comparison of average relative times $\bar{T}_p^{\mathcal{S}c1}$ in Kitchen 1 and $\bar{T}_p^{\mathcal{S}c2}$ in Kitchen 2 for each subject.

3.3.2 Time comparison

We compared the relative average time $\bar{T}_p^m$ of each subject in the scenarios performed in Kitchen 1 (ct-1-1 – ct-1-6) and those in Kitchen 2 (ct-2-1, ct-2-2). Figure 16 to visualises the relative duration in both kitchens to show if the efficency of a subject depends on a specific environment. The grey line indicates the values where the relative efficiency in both kitchens would be equal.

The maximum difference for one participant compared to its own average time ($|\bar{T}_p^{\mathcal{S}c1} - \bar{T}_p^{\mathcal{S}c2}|/\bar{T}_p^{\mathcal{S}}$) is 16%, the average is 8%. Overall, subjects who performed tasks very efficiently in one kitchen, were also among the faster performers in the second kitchen and vice versa.

Table 5: Qualitative categorisation of navigation behaviour over all trials in Kitchen 1 and scenario ct-2-2 in Kitchen 2. The table shows the number of subjects falling in each of the categories.

Kitchen 2 \ Kitchen 1	C1	C2	C3
C1			1
C2		3	
C3	2	1	2

C1: direct way
C2: long way
C3: mixed behaviour

3.3.3 Navigation paths

We also examined the paths on which the subjects navigated the robot to fulfil tasks. In Kitchen 1, an object could be placed on the far side of the table either by reaching over the table (quite a challenge with the robot's capabilities) or moving around the table (which takes longer). Similar options can be identified in Kitchen 2 as shown in Figure 9.

Table 5 shows a qualitative analysis of the paths in each kitchen using the categories defined on page 12. We show the aggregated results over all trials in each kitchen (for Kitchen 2 only scenario ct-2-2 was evaluated). In the aggregated case, category 3 also contains those participants who showed different behaviour in different scenarios in Kitchen 1.

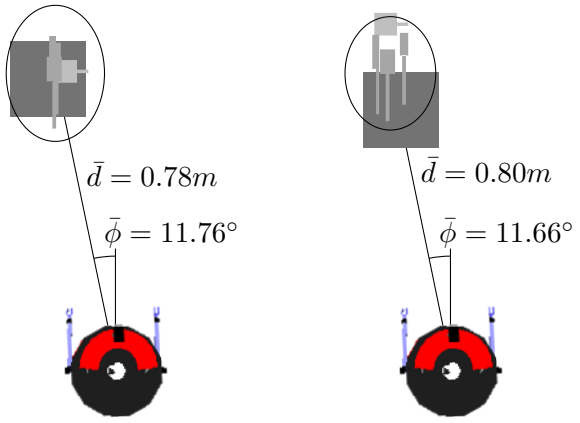
Table 5 suggests some general tendencies of some subjects to choose the long way in both kitchens (three participants). Five participants seem to avoid long ways, using the direct way or showing a mixed behaviour. Only one subject who chose long navigation paths in Kitchen 1 seemed to seek more efficient paths in Kitchen 2. The possibility of the simulation to define arbitrary worlds could be very beneficial to study human navigation behaviour by abstracting from personal physical skills (like size and agility) and concentrating on the cognitive aspects and the configuration of different worlds.

3.3.4 Gripping parameters

Finally, we compared the parameters of gripping tasks in both kitchens by measuring the robot's relative position to the objects. Figures 17a and 17b illustrate the average values of the position parameters (distance d and absolute angle ϕ) for each object type and Figures 17c and 17d show the average values and standard deviations for the trials in both kitchens.

The average values over all object types are almost identical in both kitchens: the average distance differs by 2 cm and the angle by 0.1°. Likewise, the variances are very similar. Figure 17 suggests that in Kitchen 1 all object types were gripped from about the same distance, whereas in Kitchen 2 there are visible differences in the position from which the objects were gripped. This might be due to the different topologies of the kitchens or the executed tasks. Another likely explanation is the lower number of samples in Kitchen 2, where accidental differences have a higher influence on the average value than in Kitchen 1.

Overall, the parameterization of gripping actions is comparable in the two kitchens. An interesting question would be how far the standard parameterization is changed when objects in hard-reachable positions need to be gripped.



(a) Kitchen 1. (b) Kitchen 2.

	average $\bar{d}$	st. dev. $\hat{d}$
Kitchen 1	0.78	0.18
Kitchen 2	0.80	0.15

(c) Distance.

	average $\bar{\phi}$	st. dev. $\hat{\phi}$
Kitchen 1	11.76	9.10
Kitchen 2	11.66	8.35

(d) Angle.

Figure 17: Illustration of average gripping parameters in both kitchens. The average over angles works with average values, which means that the object can also be at the right side of the robot. The ellipse around the objects indicates the standard deviation of the distance and angles. For clarity, the scaling of robot and objects are different.

4 Discussion and Conclusions

The user study has provided us with valuable observations on the usability of our current implementation of the simulation control. Our subsequent enhancement of this module has concentrated primarily on reducing failures by adding a warning mechanism when the robot is too near a piece of furniture when the user asks it to grasp an object. Because many failures are caused by problems in the underlying physics engine, this prediction will never be completely reliable, but might reduce the failure rate.

Even though the subjects of this study were no passionate computer gamers, several of them criticised the control of the robot not following the industry standard, where an explicit stop of the movement is not necessary and the character moves only as long as the respective arrow keys are pressed. We have by now changed the control mode and will verify its acceptance in subsequent studies.

Another aspect criticised by the users was the speed of the robot. The simulation can be accelerated to some extent, but this is restricted by computing power. Besides, we could make the grasping faster by violating some constraints of the arm joints. However, we cannot change these restrictions for an autonomous robot, only the manually controlled one. To keep the benefit of the similar time scale of the two robots, we will only accelerate the hand-controlled robot slightly and observe in further experiments if this affects human-robot collaboration.

The observations of the user study also indicate that more information and better training before starting the trials would help the users to avoid failures. In subsequent studies we will provide such information and prepare the users with explicit training tasks.

The results of the user study indicate that humans show similar behaviour when executing tasks in reality and simulation. The preferences of users to carry similar objects (by type or colour) at the same time is almost the same in simulation and reality. The difference that users in reality tend to prefer the carrying of objects of the same type in contrast to the more pronounced colour preference in simulation, is not a problem for the kind of research we are

interested in. We develop adaptive technology, so that a robot can observe human behaviour, derive general concepts and adapt to specific users. Even though the concepts observed in simulation are not exactly the same as in reality, the learning technology can be developed in sumulation to be used in reality to acquire the specific models there. The emphasis of our research is the development of general techniques, not specific models.

Also the subjects' preferences for placing objects when the goal positions are not predefined, show a similar picture in simulation and reality. Again, the simulated results show somewhat weaker preferences (using two colours for two place settings vs. complete sorting of colours), they are sufficiently similar for a robot to develop adaptive behaviour.

The relative time scales of performing table setting and clearing tasks are on the same level and show similar variances in simulation and real-world execution. The similar timing ensures to some extent the generalisability of simulated results, although it is doubtful that the same scaling can be observed on an action level.

The subjective evaluation of the users confirms the observed differences in preferences for handling objects. However, the subjects seem not to be aware that they show similar patterns in simulation. The results indicate that the users are more aware of their actions in simulation than in reality. For experiments, where unconscious behaviour is of importance, one might add extra stress factors to the experiment, for example by initiating a competition as in real computer games or by providing a second task, which the subjects should complete while performing the simulated household tasks.

The few trials in the second simulated kitchen indicate that the usability of the simulation depends more on the individual subject than on the world, which means that behaviour differences observed in different kitchen environments are due to the environments rather than different control capabilities. These trials have also shown that differences in higher-level behavior result more from the different kitchen environment than the differences observed between real-world execution and simulation in the same kitchen set-up.

With this user study, we could show that a simulator with a manually controlled character is a useful testbed for developing and testing human-robot interaction. It follows the idea of the "Oz of Wizard" approach [6], which claims to focus not only on human behaviour, but also to allow robot development by modelling the human. A similar approach to our simulation is the restaurant game [5] developed at MIT Media Lab and used for research on high-level interaction patterns. The restaurant game concentrates on high-level cognitive capabilities such as planning and communication, but without the underlying embodied hardware. In contrast, as we are interested in the development of complete robot systems, we don't want to abstract from the underlying physics.

Overall, this user study has provided insight into behaviour patterns for table setting and clearing tasks and how these patterns appear in a non-embodied environment such as our simulation. The results show the feasibility of studying human-robot interaction in simulation, allowing the development of realistic robot behaviour while at the same time being able to interact with humans. Moreover, such a simulation environment can be an interesting testbed for studies on cognitive behaviour, because it allows to separate the physical embodiment from pure cognitive behaviour. In addition, experiments in simulation allow the execution of arbitrary worlds at low cost and facilitate the data collection significantly.

Acknowledgements

I thank Anna Schubö and Tamara Lorenz for their advice in setting up the user study. Thanks are also due to all the participants of the study and the team in the robot kitchen for making it available for the real-world trials. This work was supported by DFG excellence initiative research cluster Cognition for Technical Systems.

References

[1] Michael Beetz, Jan Bandouch, Alexandra Kirsch, Alexis Maldonado, Armin Müller, and Radu Bogdan Rusu. The assistive kitchen — a demonstration scenario for cognitive technical systems. In *Proceedings of the 4th COE Workshop on Human Adaptive Mechatronics (HAM)*, 2007.

[2] J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41, 1984.

[3] Manja Lohse, Marc Hanheide, Katharina J. Rohlfing, and Gerhard Sagerer. Systemic interaction analysis (sina) in hri. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009.

[4] Armin Müller and Michael Beetz. Designing and implementing a plan library for a simulated household robot. In Michael Beetz, Kanna Rajan, Michael Thielscher, and Radu Bogdan Rusu, editors, *Cognitive Robotics: Papers from the AAAI Workshop*, Technical Report WS-06-03, pages 119–128, Menlo Park, California, 2006. American Association for Artificial Intelligence.

[5] Jeff Orkin and Deb Roy. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development (JOGD)*, 3(1):39–60, December 2007.

[6] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. The oz of wizard: Simulating the human for interaction research. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009.

[7] Ryohei Ueda, Takashi Ogura, Kei Okada, and Masayuki Inaba. Design and implementation of humanoid programming system powered by deformable objects simulation. In *Proceedings of the 10th International Conference on Intelligent Autonomous Systems*, pages 374–381, 2008.

[8] Luke S. Zettlemoyer, Hanna M. Pasula, and Leslie Pack Kaelbling. Learning planning rules in noisy stochastic worlds. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI)*, 2005.